ベイズ統計と統計物理 : 有限温度での情報処理

AUTHOR(S):

伊庭, 幸人

CITATION:

伊庭, 幸人. ベイズ統計と統計物理 : 有限温度での情報処理. 物性研究 1993, 60(6): 677-699

ISSUE DATE:

1993-09-20

URL:

http://hdl.handle.net/2433/95176

物性研究　60－6（1993－9）

ベイズ統計と統計物理

— 有限温度での情報処理 —

伊庭幸人

統計数理研究所
〒106 東京都 港区 南麻布 4-6-7
Email: iba@ism.ac.jp

（1993年11月15日受理）

ベイズ統計の“有限温度の情報処理”としての側面を物理の研究者向けに解説し、関連する自分の研究を紹介した。とくに、

1. ベイズ統計の枠組と統計物理との類似。
2. ‘有限温度’での推定の最良性。
3. ‘自由エネルギー’の情報処理における重要性。
4. 有限温度での推定を行なうためのアルゴリズム。

について述べた。付録では、spin glass における Nishimori line とベイズ統計の関連について、hyper parameter の推定という観点から考察した。

KEYWORDS
ベイズ統計, 事後分布、損失関数、optimal estimator、hyper parameter、marginal likelihood、自由エネルギー、multicanonical algorithm、Metropolis-coupled chain、Nishimori line

1　ベイズ統計

1.1　枠組

まずベイズ統計の枠組について簡単に述べる。特定のパラメータの値 x に対して、データ y が出現する確率を

$$L(y|x) = \frac{\exp(-E_L(y|x))}{Z_L} \tag{1}$$

$$Z_L = \sum_y \exp(-E_L(y|x)) \tag{2}$$

とする。ここで $\sum_y$ はあらゆる y についての和である。これを、x の関数とみるとき、likelihood(ゆう度)と呼ぶ。以下では、Z_L は x を含まないとする。

y を知ったときに x を推測するためには、x についての事前知識が必要である (とベイズの立場では考える)。これを確率分布の形に表したものが事前分布 (prior distribution) である。

$$\pi(x) = \frac{\exp(-E_\pi(x))}{Z_\pi} \tag{3}$$

$$Z_\pi = \sum_{config.} \exp(-E_\pi(x)) \tag{4}$$

ここで、$\sum_{config.}$ はあらゆる x についての和である。

すると、y を得たときの x の確率は、簡単な計算により、以下の分布 (事後分布、posterior distribution) で表せる (Bayes theorem)。

$$P(x) = \frac{L(y|x)\pi(x)}{\sum_{config.} L(y|x)\pi(x)} = \frac{\exp(-E_{pos}(x))}{Z_{pos}} \tag{5}$$

$$E_{pos}(x) = E_L(y|x) + E_\pi(x) \tag{6}$$

$$Z_{pos} = \sum_{config.} \exp(-E_{pos}(x)) \tag{7}$$

以上では、統計物理を連想させるような書き方 ($\exp(-E_{pos}(x))$ など) をわざと用いた。これが、実質をともなうかどうかは場合によるが、x が非常に大きな次元のベクトルである場合には、アナロジーが成り立つ。このような例としては、画像再構成、曖昧さを許す分類、畳み込み符号による通信などが考えられる。最近では、統計の文献でも、統計物理風の表示 (Gibbs 分布による表示) がかなり使われるようになっている。また、$\pi(x)$ の部分については、できるだけ偏見のないことを表す分布 (ignorant prior) を用いるのが普通であったが、あとの例に見るように、積極的に主観をいれてそれを学習によって調節するという立場 (informative prior を用いた経験ベイズ法) が注目されてきている。より詳しい解説は、(Iba, unpublished note[0])) に書いた。これは、近い将来に出版する計画である。

1.2　例

- 例: ガウス型の平滑化

$$x_i \in R \tag{8}$$

$$E_L(y|x) = \alpha \sum_i (x_i - y_i)^2 \tag{9}$$

(1階)

$$E_\pi(x) = \gamma \sum_i (x_{i+1} - x_i)^2 \tag{10}$$

(2階)

$$E_\pi(x) = \gamma \sum_i (x_{i+1} - 2x_i + x_{i-1})^2 \tag{11}$$

- 例: Caucy 分布による 平滑化・変化点検出 (Kitagawa(1987))

$$x_i \in R \tag{12}$$

$$E_L(y|x) = \alpha \sum_i (x_i - y_i)^2 \tag{13}$$

(1階)

$$E_\pi(x) = \sum_i log(\tau + (x_{i+1} - x_i)^2) \tag{14}$$

- 例: イジング模型 (一般にはポッツ模型) による画像再構成 (Geman and Geman(1984))

$$x_i \in \{\pm 1\} \tag{15}$$

(Binary Symmetric Channel)

$$E_L(y|x) = -\alpha \sum_i x_i y_i \tag{16}$$

(ガウス雑音)

$$E_L(y|x) = \alpha \sum_i (x_i - y_i)^2 \tag{17}$$

$$E_\pi(x) = -J \sum_{(i,j)} x_i x_j \tag{18}$$

- 例: 曖昧さを含む分類 (ここでは、2つに分ける場合)

以下で、f、g は問題の詳細によって決まる関数。

$$x_i \in \{\pm 1\} \tag{19}$$

(各 x_i について情報がある場合: 2-mixture)

$$E_L(y|x) = -\sum_i f(y_i) x_i \tag{20}$$

(相互的な情報がある場合)

$$E_L(y|x) = -\sum_i f(y_i) x_i - \sum_{(i,j)} g(y_{ij}) x_i x_j \tag{21}$$

(2 グループの大きさが同じくらいと仮定)

$$E_\pi(x) = const. \tag{22}$$

(2 グループの大きさが $\exp(h)$ 対 $\exp(-h)$ くらいと仮定)

$$E_\pi(x) = -h\sum x_i \tag{23}$$

(大きさが'わからない'ということのひとつの表現)

$$E_\pi(x) = -\log\frac{n!m!}{(n+m)!} \tag{24}$$

$$n = \sum_i \frac{1+x_i}{2} \tag{25}$$

$$m = \sum_i \frac{1-x_i}{2} \tag{26}$$

1.3　ベイズ統計の "有限温度" 性

統計物理とのアナロジーでいえば、ベイズ統計は本質的に "有限温度" 的である。このことは、"解" が事後分布 $P(x)$ という "分布" であることに現れている。もし、温度 T での分布を、

$$P_T(x) = \frac{\exp(-E_{pos}(x)/T)}{Z_T} \tag{27}$$

$$Z_T = \sum_{config.} \exp(-E_{pos}(x)/T) \tag{28}$$

と定義すれば、これは、温度 $T = 1$ で事後分布に一致し、温度 $T \to 0$ で一点 (事後分布の global maximum) に収縮することになる。最適化で求められるのは後者であって、これはベイズの立場からすれば、事後分布に含まれている情報の一部を取り出したにすぎない。逆にいえば、ベイズの立場をとることで、"local maxima の回避" などに帰着できない、有限温度の情報処理ということが見えてくる可能性がある (Iba(1989))。以下では、そのいろいろな側面を見ていくことにする。

2　ベイズ統計では推定値は損失関数によって違う

"解が分布である" といっても、結局はひとつの "推定値" を出さなければならないことも多い。この場合、推定の目的がなにかによって、推定値が違ってくる。形式的には、"目的" を損失関数 (loss function) によって与えることで、それに対して最適な推定値を与える関数 (optimal estimator) が定まり、それを適用することで、事後分布から推定値 (estimate) が得られることになる。推定値の "信頼区間" や "誤り確率" も事後分布から得られるし、それを使って次のデータを獲得する計画を最適化することも可能であるが、それらの側面については、ここでは触れない。

2.1　例: 平滑化・変化点検出のモデルの場合

たとえば、次のような推定値の定義が可能である。これらがすべて同じなのは、事後分布がガウス分布のときのみである。

1. 事後分布全体の global maximum (MAP 推定値 = Maximum A Posteriori Estimate) を $\{x_i\}$ の推定値とした場合。
2. i ごとに x_i の周辺分布 $P_i(x_i)$ を考えてその global maximum を x_i の推定値とした場合。ただし、

$$P_i(x_i) = \int \Pi_{j\neq i} dx_j P(\{x_i\}) \tag{29}$$

3. x_i の事後分布による期待値を推定値とした場合。
   x_i の $P_i(x_i)$ での期待値といっても同じである。

4. $P_i(x_i)$ のメジアン (中央値) を推定値とした場合。

(1),(2),(3),(4) にそれぞれ対応する損失関数はそれぞれ次のようになる。$\{x_i^{true}\}$ を真の値、δ を十分小さい数とする。

1. "任意の i について $|x_i^{true} - x_i| < \delta$" である確率を最大にする推定値。

2. "$|x_i^{true} - x_i| < \delta$ である i の数の期待値" を最大にする推定値。

3. "$\sum_i |x_i^{true} - x_i|^2$" を最小にする推定値。

4. "$\sum_i |x_i^{true} - x_i|$" を最小にする推定値。

計算のやり方から見ると、(1) を 温度零 の推定値、(2),(3),(4) を 温度 1 の推定値と見ることもできる。annealing やその他の最適化手法が求めようとするのは (1) である。

2.2 歴史

- 推定値が損失関数によって違うことは古くから知られている。モデルの大規模化 (x の次元が大きくなる)、informative prior の導入、非ガウス化、統計物理のアルゴリズムの導入などでどう情勢が変わるかが問題である。

- 画像処理について、マルコフ鎖モンテカルロ法 (*) との関連でこの点を強調したのは、Marroquin(1985) である (これは preprint で、論文としては Marroquin et al.(1987) の一部に組み込まれている)。

  しかし、その論文にも既知のこととして引用文献が示されている。そのほか、たとえば、Derin et al.(1984)、Kay and Titterington(1986)、Devijver(1987)、Zhang(1992) などにも、2 値画像について '真の画像との重なり' を最大にする Bayesian optimal estimator が用いられている。

  (*) メトロポリス法や Gibbs Sampler の総称。物理以外の分野では単にモンテカルロ法といったのでは非常に広い範囲を指すので、区別した名称が必要である。筆者は 'メトロポリス的モンテカルロ法' という名称を使ってきたが、最近では 'マルコフ鎖モンテカルロ法' に統一する動きがあるようなので、それに従う。

- 筆者は数年前にこの問題に気づいたが、2 値画像については Marroquin の論文があることを知ったので、別の問題 ("3 すくみの分類の問題") で別の効用関数についてこの点を調べて、メトロポリス法の使い方の注意ということで発表した。しかし、のんびりやっているうちに、マルコフ鎖モンテカルロ法を温度 1 で (=annealing でなしに) 使うことは統計の分野では周知の事実となっていたようで、論文は落されてしまった (Iba(1992))。

- 最近、Ruján という人が Phys.Rev.Lett. に "Finite temperature error correcting codes" という題目で、Marroquin などが画像に対して述べたこととほぼ同じことを符号解読について述べた論文を書いている (Ruján(1993))。それに対して、それを証明したという follow も出ている (Nishimori(1993)、Soulas(preprint))。符号解読の場合は、ある種の gauge 対称性をもつクラスが自然であるなどの特別な事情があるが、有限温度での最良性自体は新しいことではない。物性物理への feedback があれば面白いかもしれないが、なかなか難しいだろう (付録も参照)。本稿との温度の定義の違いに注意されたい。

  筆者は、本稿の主題とは別な観点からも、符号理論とランダム系の統計物理の関係を考察したことがある (Iba, unpublished note[3])。今回の話題はそちらの方とはあまり関係がなさそうである。

2.3　いままで研究した例

損失関数を決めれば、特定の estimator が良いということは比較的簡単に示すことができる。したがって、考えられる問題の設定としては

1. simulation を行なって、特定の損失関数についての推定値と事後分布全体の global maximum にもとづく推定値の差が非常に大きい場合があることを示せるか?
2. 実際のデータに対してはどうか? この場合、損失関数の選び方が実感と合っているかという問題も同時に考えなくてはならない。
3. '温度1での推定値' を得るためのアルゴリズムは?

などがある。(3) についてはあと (5節) で扱う。

一般に、推定が非常に容易な条件 (データが多い、雑音が少ないなど) のもとでは、どのような reasonable な損失関数に対しても、似たような推定値が得られると思われる。また、推定が非常に困難な場合は、どのような方法をとってもうまくいかない。また、次節 (3節) で扱うように、likelihood や事前分布が正確に分かっていることはまれであって、そのなかに未知パラメータ (hyper parameter) が含まれていることが多いが、それをデータ自体から推定することは悪条件のもとでは困難である。このため、差がでるような条件は限られているものと思われる。さらにきびしい意見としては、条件が悪くなるにつれて結果が滑らかに悪化していくような<u>モデル</u>を作ることが重要で、分布の広がりなどに頼るのは邪道だという考えもある。

何人かの人が、いろいろな状況で損失関数による違いを調べているが、あまり差がないという結果が多いようである (たとえば、Geman and McClure(1987))。筆者が調べた例としては次のようなものがある。中にはかなり差のあるように見えるものもあるが、結局どの例に対しても決定的なことはいえていない。

- イジング模型による画像処理の場合

  イジング模型を事前分布に使った画像処理については、前に述べた通り Marroquin が、真の画像との重なりが最大になるような推定値と MAP 推定値に大差があると主張している。しかし、そこにあげられた例は誤りと思われる (模擬データを作るのに走らせたモンテカルロ計算が緩和していない)。しかし、正しい例でやった結果 (Iba, unpubished note[1]) でも、差はでる。hyper parameter((18) の J) をデータから推定することを要求すると、より苦しくなるが、大きな差がでる場合もある。hyper parameter の推定を中心にしてまとめた論文 (Iba(1991a)) では、"重なり最大" の推定値を使ったが、MAP 推定値との差に関することは詳しく書かなかった。イジング模型を事前分布とした場合、温度を下げると相関が発達しすぎることは物理的にあきらかだから、この場合に温度1と零で差がでることはわかりやすい。しかし、イジング模型を事前分布に使った画像処理が本当に良いものかどうか、その場合の実際の '使われ方' がベイズ的解釈にふさわしいものか、多くの疑問が残されている。

- "曖昧さを含む分類" の問題の場合

  "曖昧さを含む分類" というのは、筆者が勝手に名付けた問題で、適当な情報に基づいて、要素をいくつかのグループに分ける問題をさす。もっとも簡単な場合は、事後分布において、パラメータの間に相互作用のない場合 (20) である。この場合、問題となるのは、関数 f に含まれている hyper parameter の推定である。このタイプのモデルは finite mixture としてよく知られている。これをもっと難しくしたのが、(21) のような問題で、graph-bipartitioning の有限温度版とも見られる (graph-bipartitioning は最適分割を求める問題であるのに対し、ここで考えているのは、分割を<u>推定する</u>問題であることに注意)。Rujàn らの扱っている伝送路も基本的には、このモデルに近いものである。ただし、伝送路の場合は、gauge invariance を仮定しても不自然にならない点が異なる。

Iba(1992) では、"曖昧さを含む分類" の問題に対し、温度 1 でメトロポリス法を適用し、それによって、optimal estimator を計算できることを示した。この例では、分類のもとになる情報として勝敗表を使ったので、'i が j に勝つ確率' の optimal estimator について論じた。KL 情報量を損失関数とした場合、温度 1 でベストの結果が得られ、かつ温度零よりはるかに良いことが示されたが、これは KL 情報量の特性 (よくわからないときは確率 1/2 と答えた方が有利になる) によるところが大きい。

Iba(1992) では (22) の事前分布を用いたが、分類結果のグループサイズが著しく不均一な場合のことを考えると、(24) の事前分布を用いるか、(23) を用いて h を 3節の方法でデータから推定するのが望ましい (**)。この場合、強い '強磁性' の相互作用を導入することになるので、結果がいろいろ変わってくる。

(**) この 2 つの方法は、一見ずいぶん違って見えるが、実際は非常に近い (たとえば、(Iba(1991c) に解説がある)。

- Caucy 分布による平滑化・変化点検出の場合

すでに統計的情報処理としての有効性があきらかにされている問題に関して実験することが重要なように思われる。そこで、'複雑系 2' の発表では、Caucy 分布による平滑化・変化点検出 ((13,14)) の場合に、損失関数によってどう推定値が変わるかを調べて発表した。Kitagawa(1987) には、非線形フィルター (転送積分法、5.1節参照) を用いて計算した周辺分布 $P_i(x_i)$ とそのメジアンが示されているが、温度零での推定値 (MAP 推定値) との比較は行なわれていない。結果は図 1a-e、図 2a-e に示す。

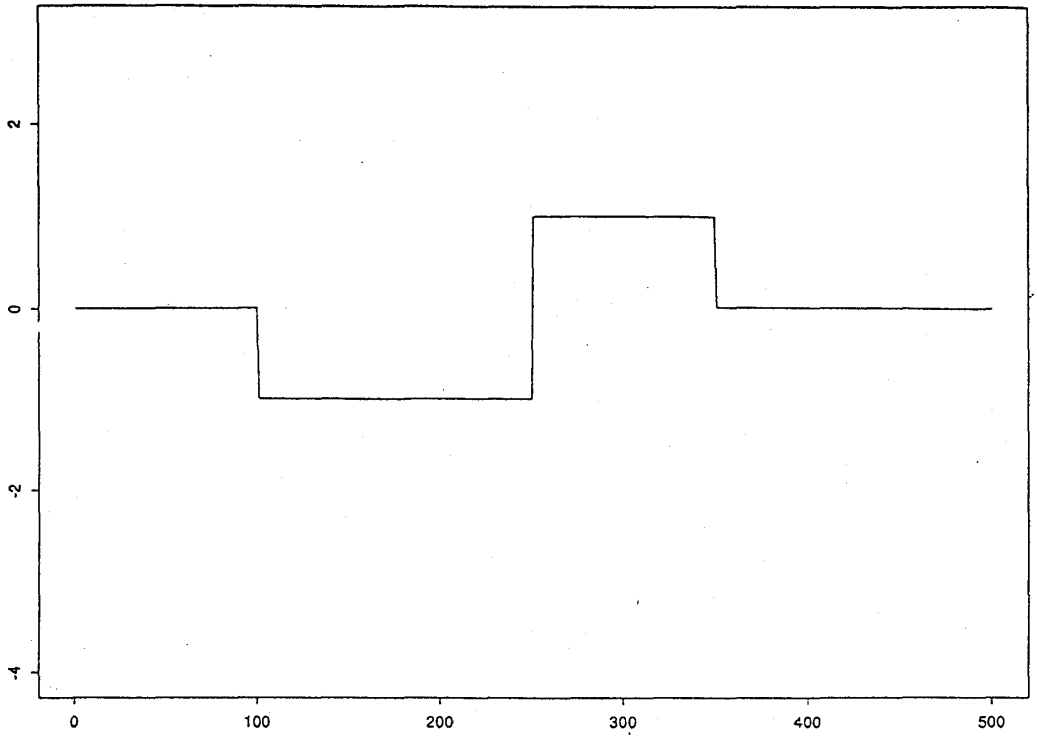

図 1 a　真の値

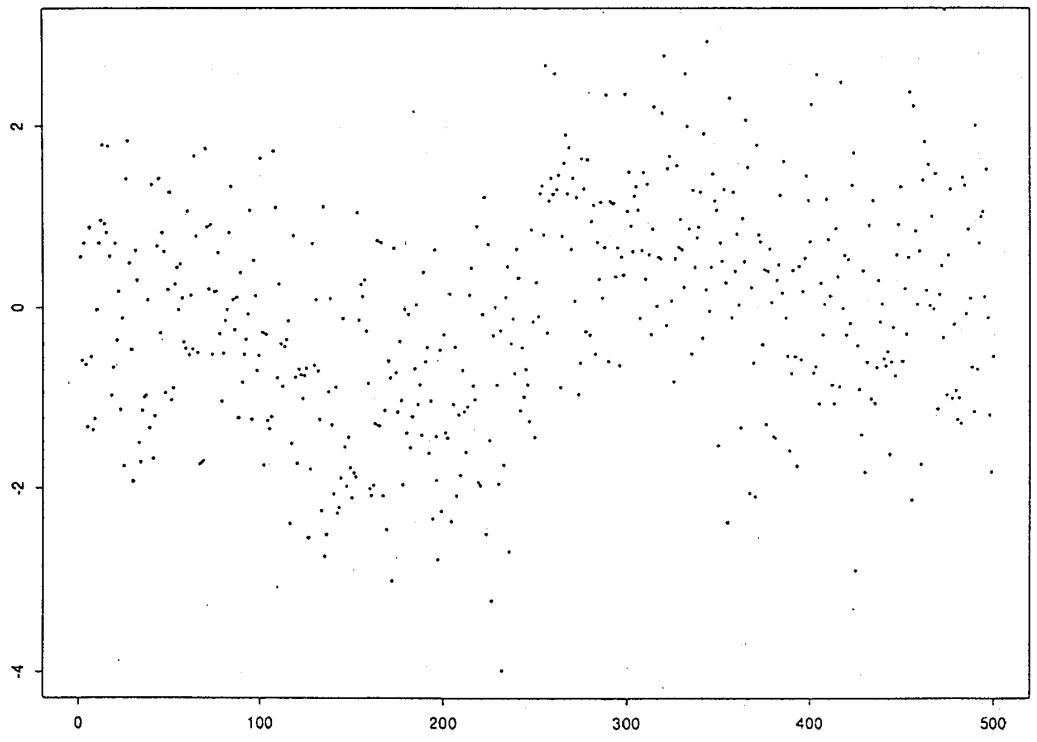

図１ｂ　データ

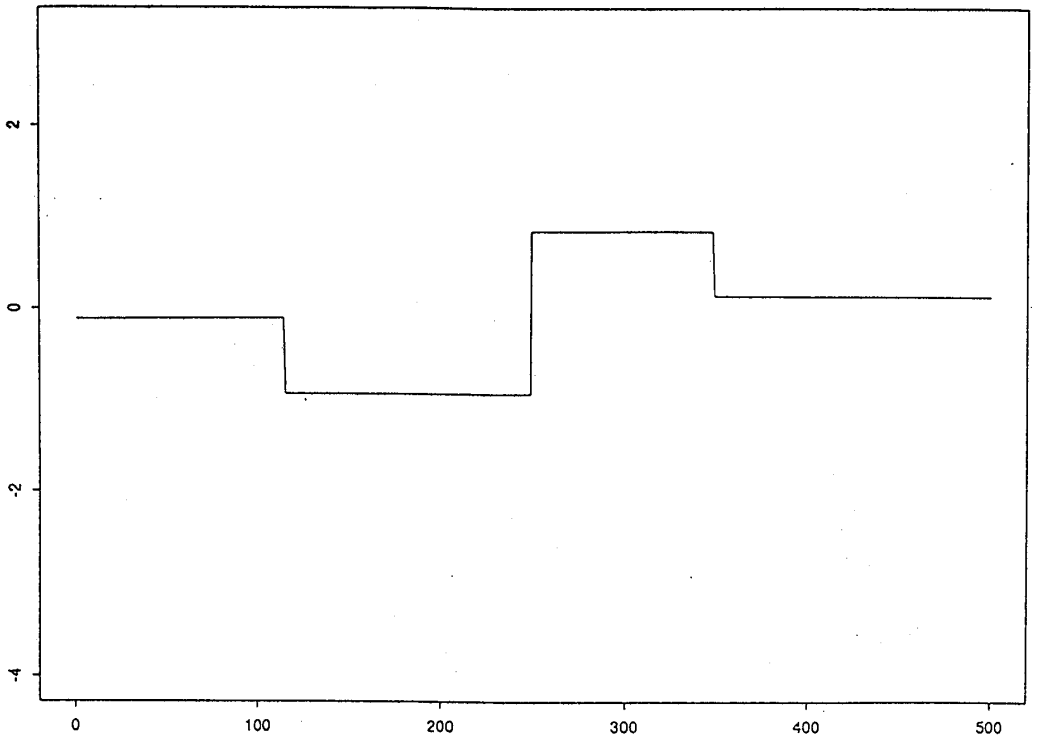

図１ｃ　事後分布の最大値（温度 零）

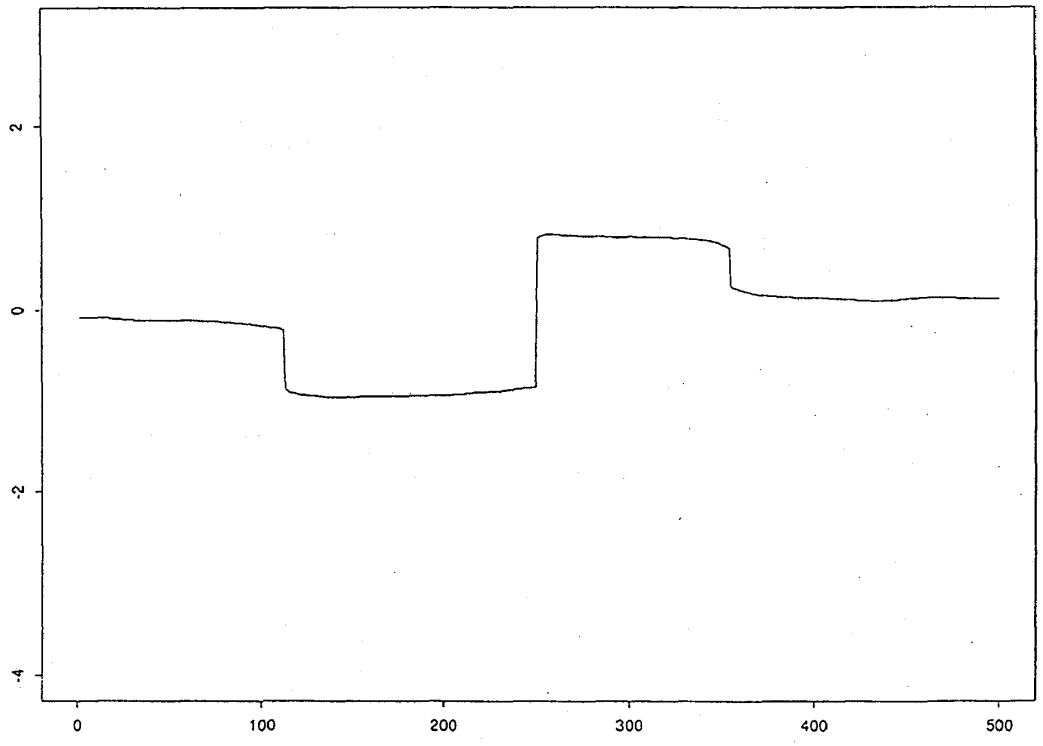

図１ｄ　周辺分布 $P_t(x_t)$ の最大値

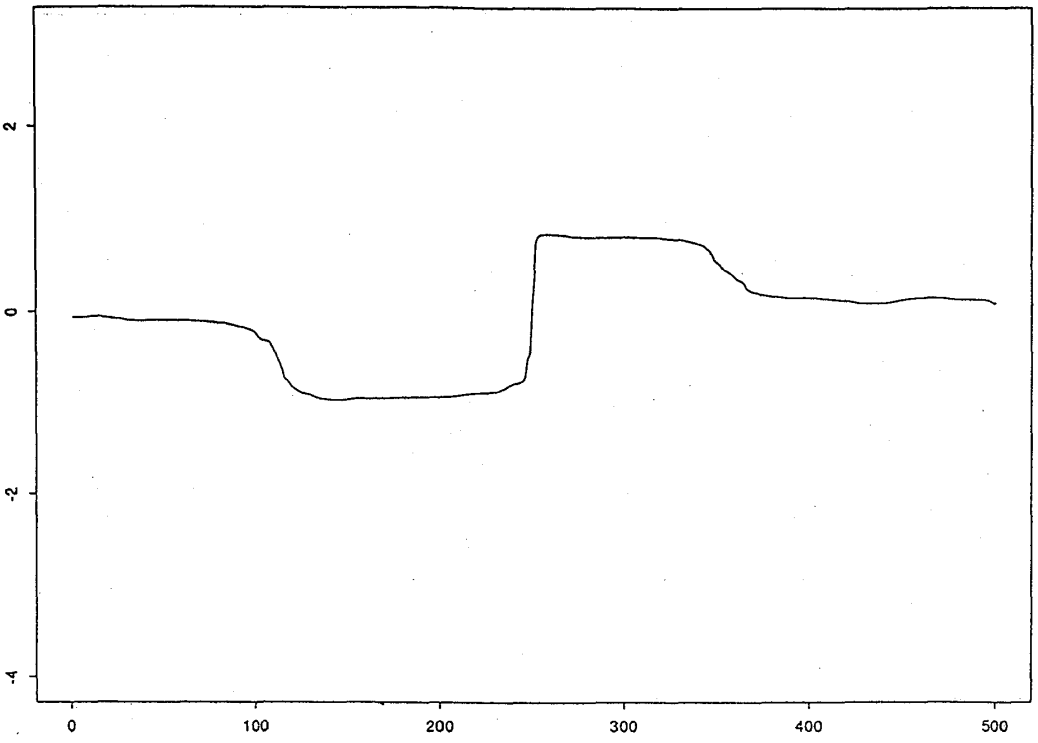

図１ｅ　期待値

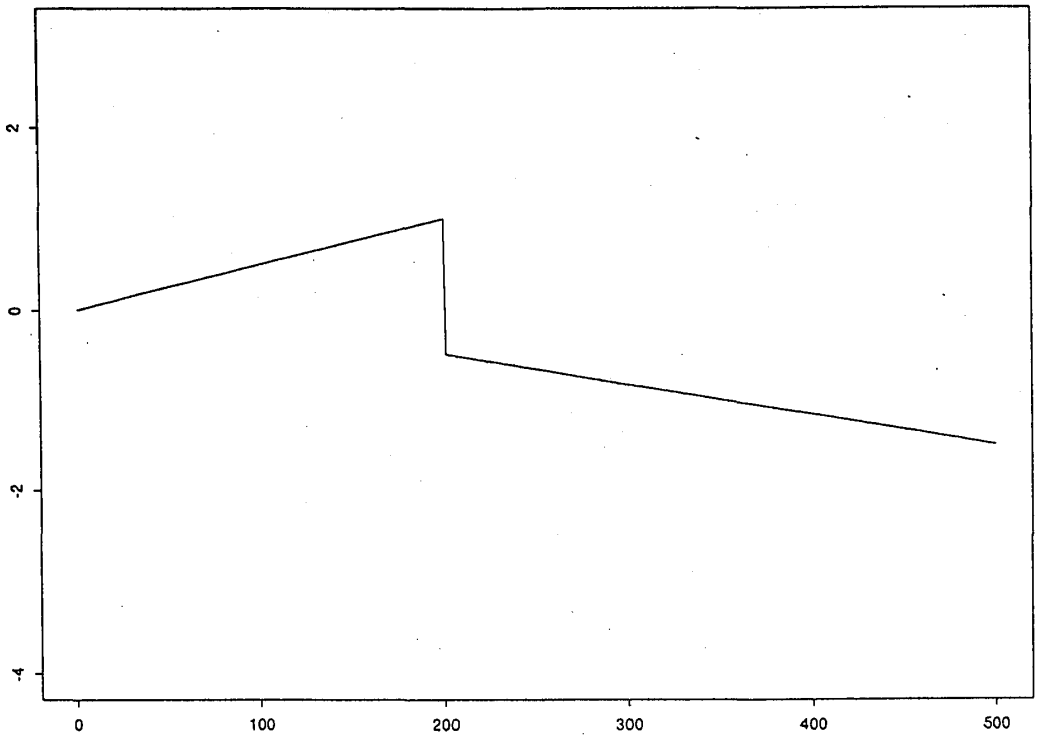

図２ａ　真の値

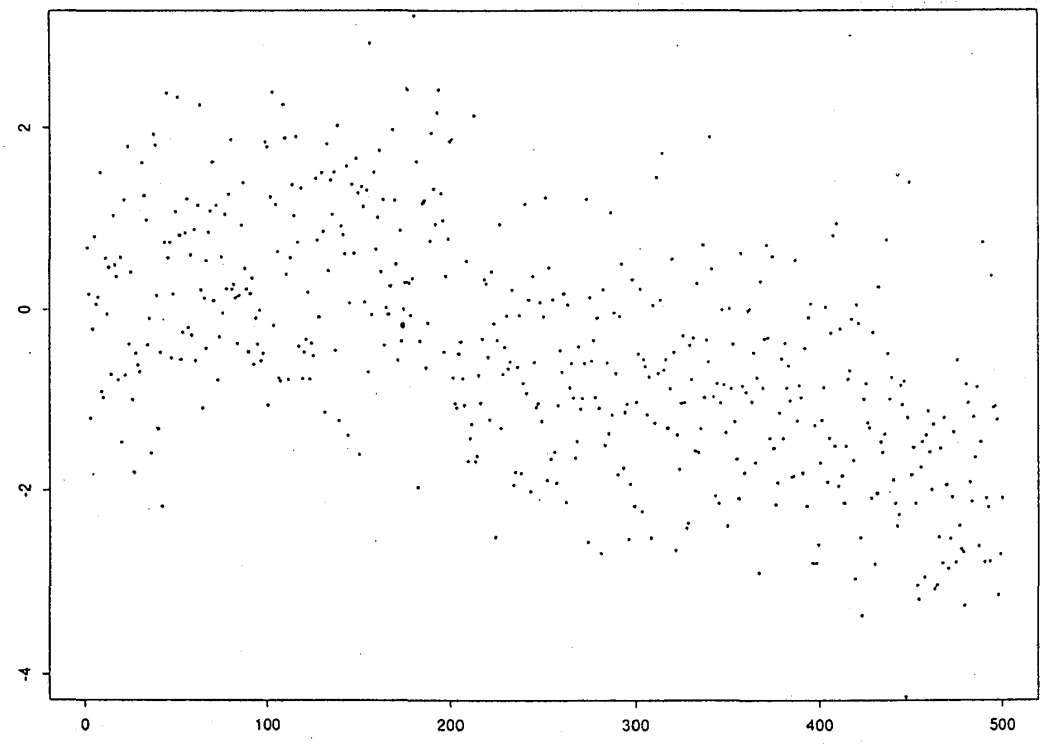

図２ｂ　データ

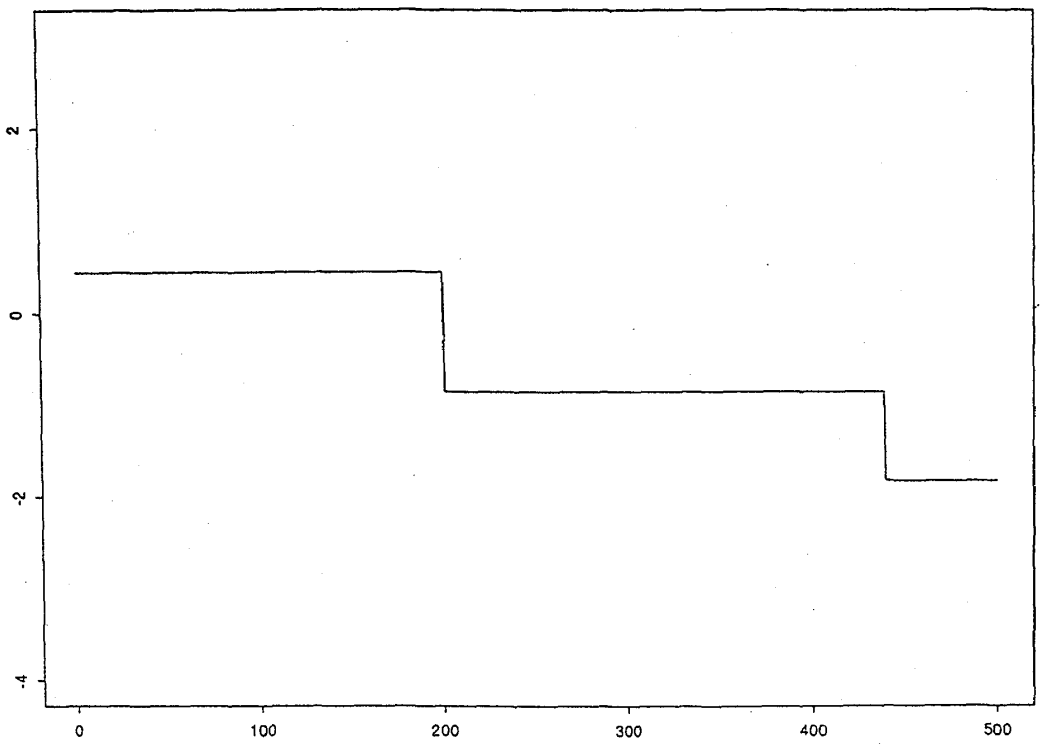

図２ｃ　事後分布の最大値（温度 零）

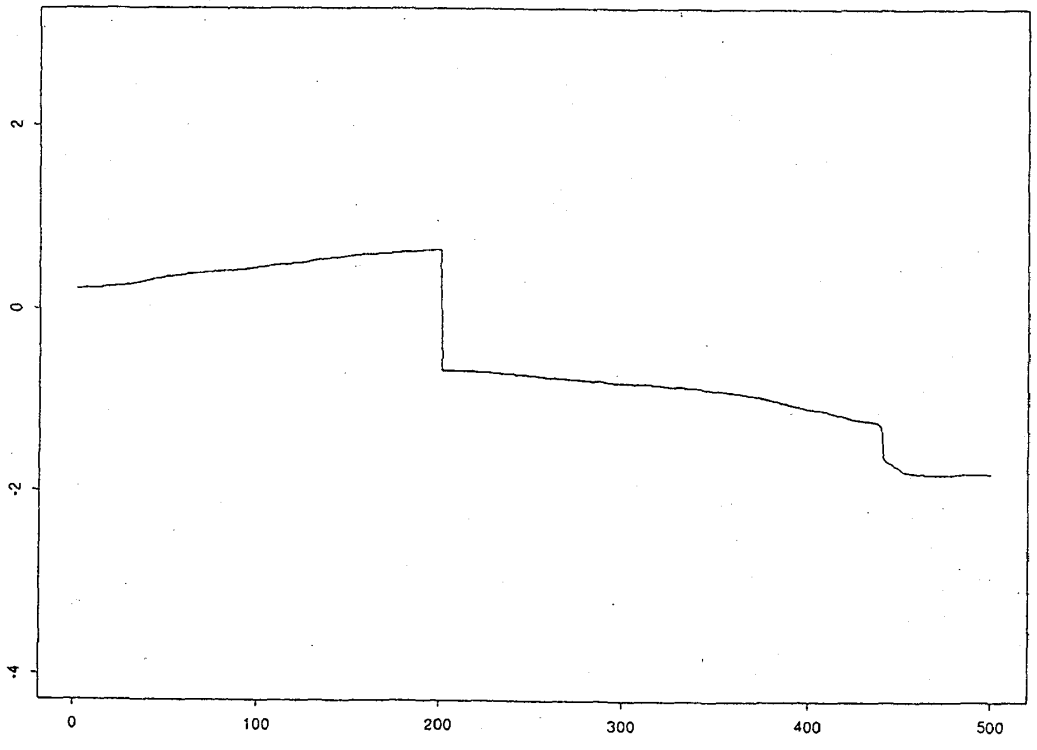

図２ｄ　周辺分布 $P_t(x_t)$ の最大値

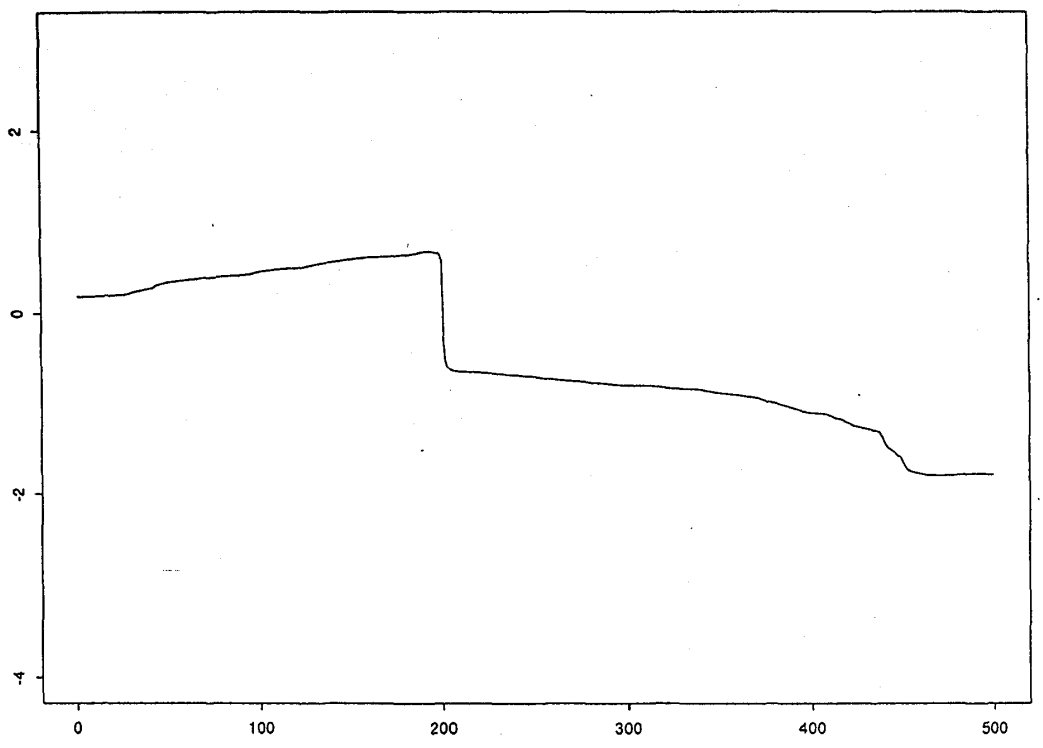

図 2 e　期待値

図 1a 及び図 2a のデータに対してガウス雑音を加えて作ったデータが図 1b 及び図 2b である。ガウス雑音の分散は 1.0 である。図 1b 及び図 2b のデータに対して非線形フィルターを用いて、推定値の計算を行なった結果を図 1c-e、図 2c-e に示す。2.1節の (1),(2),(3) に対応する結果が、それぞれ (図 1c、図 2c),(図 1d、図 2d)、(図 1e、図 2e) に示されている。hyper parameter τ, α は、Kitagawa に従って、marginal likelihood を最大化する方法 (次節 (3節) 参照) を用いてデータから推定した。$\tau = 10^{-5}$、α は図 1c-e では 0.45、図 2c-e では 0.5 である。

図 1 の例では、図 1c に示した温度零の推定値 (MAP 推定値) が直観的にはもっとも優れているように思われる。これに対して、図 2 の例では逆に図 2c はおかしく感じられ、図 2d や図 2e の方が自然である。図 2 の場合には、温度零と温度 1 の違いは非常に大きい。

ただし、これらの場合に (14) から導かれる分布はきわめて尖っており、ここで行なった数値計算 (とくに温度零の場合) が正しい結果を与えているかどうかは問題がある。図 1、2 は暫定的な結果と考えてほしい。また、ここで見られる違いが数値計算の artifact でないとしても、それが本当に有意義なものかどうかは、さらに検討すべき点がいくつかある。

3　hyper parameter の推定

hyper parameter という用語はすでに使ったが、ここであらためて説明する。hyper parameter(超パラメータ) とは、事前分布や likelihood に入っている 1 つレベルの高いパラメータを意味する。応用領域や

著者によっては、パラメータのことを state vector とか pixel(画素) とか呼んで、hyper parameter をパラメータと呼ぶ場合もある。

$$E_L(y|x) = \alpha \sum_i (x_i - y_i)^2 \tag{30}$$

$$E_\pi(x) = \gamma \sum_i (x_{i+1} - 2x_i + x_{i-1})^2 \tag{31}$$

なら、α と γ が hyper parameter である。informative prior を使う場合には、hyper parameter の推定の問題が特に重要になる。

hyper parameter を推定する方法として、ベイズの立場から考えられるのは次の 2 つである。他に、たとえば cross validation による方法 (データを学習用のものと検証用のものに分ける方法) などもあるが、ここでは触れない。

- "エネルギー (温度零) 最小"(parameter と hyper parameter の同時最適化)

$$-\log\{L_\alpha(y|x_{MAP})\pi_\gamma(x_{MAP})\} \rightarrow \text{最小} \tag{32}$$

あるいは、

$$E^{\alpha,\gamma}_{pos}(x_{MAP}) - (-\log Z^\alpha_L - \log Z^\gamma_\pi) \rightarrow \text{最小} \tag{33}$$

- "自由エネルギー (温度 1) 最小"(maximum marginal likelihood)

$$-\log \sum_{config.} L_\alpha(y|x)\pi_\gamma(x) \rightarrow \text{最小} \tag{34}$$

あるいは、

$$-\log Z^{\alpha,\gamma}_{pos} - (-\log Z^\alpha_L - \log Z^\gamma_\pi) \rightarrow \text{最小} \tag{35}$$

ここで、x_{MAP} は事後分布を最大にする x(MAP 推定値) を、$\sum_{config.}$ はあらゆる x に関する和 (積分) を表わす。

興味深いことに、事後分布がガウス的な場合 (x が実数成分のベクトルで、$E_{pos}(x)$ が 2 次式の場合) にすら、両者は同じ結果を与えない。実際に、前者では全く駄目で、後者ではうまくいくのが普通である。これは、前者では多くの (= パラメータの数 + 超パラメータの数の) パラメータを同時に推定しているため、overfitting(過剰学習) の状態になるためだと考えられる。有限温度での情報処理という問題は、generalization の問題とも深く関係しているのである。

問題によっては、どちらの基準をとってもあまり変わらないこともあるらしい。たとえば、"分類の問題" で (23) 式の h をデータから推定する場合には、筆者の扱った例題では、どちらを使ってもあまり差はなかった。このあたりの違いについては、十分理解されているとはいい難い。

ベイズ統計のなかで閉じた問題として hyper parameter 推定の問題を扱えば、自由エネルギーを使った方が正しい hyper parameter の推定値を与えることは当たり前に近く、それほど面白いことではない (***)。両者の差が非常に大きくなることがあり、それが実際のデータ処理の局面で重要な役割を演じる、ということがポイントである。筆者の印象としては、Bayesian optimal estimator の議論よりもこちらの方が情報処理にとっても統計物理にとっても面白いのではないかと思う。

(***) gauge invariance のある場合に '自由エネルギーによる hyper parameter の推定' を適用すれば、Nishimori line について多少の洞察が得られる (付録参照)。また、SK 模型に相当する問題では、相転移と hyper parameter の推定可能性に関係があることもいえるが、いずれもややマイナーな話である。

専門的になるが、'marginal likelihood の帰無仮説下の分布' がどうなるか、という問題がある。この場合、'帰無仮説下' というのは、'データがランダム場合' と考えて良い。1 次元のガウス的な平滑化の場合にこの分布が通常のカイ 2 乗分布にならないということが、Yanagimoto and Yanagimoto(1987) に示されている。これは要するに赤外発散のせいである。赤外発散であるから、空間次元によるわけで、そのことを指摘して喜んでいた (Iba(1991b)) のだが、実は、Wahba の本 (Wahba(1990)) に本質的な部分は書

いてあった。大規模モデルの統計学のなかで空間次元が関係してくる問題には、統計物理からみて興味深いものがある可能性がある。

4　歴史

このテーマについては、一般論、実例、マルコフ鎖モンテカルロ法との関連など、多くの仕事がすでになされている。しかし、統計物理の面からの理論的接近は、筆者が無知なためかもしれないが、あまり聞かない (Zee(1987),Bialek and Zee(1987)) に 'path integral による情報処理' といった題目で関連したことがいくらか論じられている)。理論的な取り扱いの発展が望まれる。

- パラメータについて和を取った likelihood(marginal likelihood) によって、hyper parameter を推定する方法は古くから知られていたらしい。経験ベイズ法 (emphirical Bayes) ないし階層ベイズ法 (hierarchical Bayes) と呼ばれることが多い。しばしば引用されるのは Good(1965) である。ただし、昔は計算上の困難から、本格的な応用 (自由エネルギーとのアナロジーが意味をもつような応用) は難しかったと思われる。

- 近年になって、この方法は、統計のさまざまな分野で、特に大規模モデルに対して応用されるようになった。ひとつの流れとしては、赤池とその協力者によるものがある (1980 年頃から)。赤池は marginal likelihood をひとつの情報量規準とみて ABIC と呼び、marginal likelihood によって、hyper parameter を推定する方法の有効性を示した (Akaike(1980))。この仕事は、統計数理研究所の他のメンバーによって拡張され広範な領域に応用された (非常に多くの論文があるが、たとえば、Sakamoto(1985)、Tanabe(1985)、Kitagawa and Gersch(1985)、Kitagawa(1987)、Ogata(1990) やその引用文献を参照)。扱われた問題の多くは、ガウス的なものやガウス近似 ('半古典近似') が有効なものであるが、Kitagawa(1987) のように非ガウス性の強いものも含まれる。

  べつの流れとしては、たとえば、Dempster らによって普及された EM 法 (Expectation-Maximization algorithm) がある (Dempster et al.(1977))。EM 法そのものはアルゴリズムとして提示されているが、内容的には隠れた変数 (欠測値) について和をとった likelihood を最大化することにほぼ対応する。

- Hinton らの提案した、'Boltzmann machine の学習方程式' は隠れたパラメータについて和をとった likelihood を最大化する式である (Hinton and Sejnowski(1986))。これは多分、marginal likelihood の最大化にマルコフ鎖モンテカルロ法を使った最初の例であろう。ただ、温度 1 で計算する (marginal likelihood を使う) ことの意味については少ししか論じられていないため、多くの読者は単に annealing という文脈でしか理解しなかったように思われる。この例において、(1)marginal likelihood を最大化。(2)(適当に選んだ) local minimum について平均した likelihood を最大化。(3)global maximum の likelihood を最大化。の 3 つが学習においてどの程度違うのかは不明である。

  Geman らも marginal likelihood の最大化にマルコフ鎖モンテカルロ法を使っている (Geman and McClure(1987))。これは、Hinton らの仕事にヒントを得たものかもしれないが、和をとるべきパラメータの数が、hyper parameter の数よりずっと多いので、'free energy' らしくなっている ('Boltzmann machine の学習' では両方とも多い)。

- 最近では、MacKay という人が neural network での汎化 (の程度を決める) の問題に、marginal likelihood の最大化を適用して、注目されているらしいが、文献 (MacKay(1992)) はまだ入手していない。

5　"有限温度" の量をどうやって計算するのか?

5.1　有限温度のアルゴリズム

'有限温度の計算'、もっと一般に通用するような言い方をすれば、'高次元空間の上の確率分布の期待値の計算' を行なうためにはそれなりのアルゴリズムが必要である。それは統計物理の独壇場である、と

いえれば良いのであるが、実際にはそうともいえない。とくに、転送行列・転送積分に類する方法は、確率分布を扱うための方法として、工学・統計学の世界では広く用いられている。これらの分野で最も新鮮に受け止められたのは、メトロポリス法やGibbs samplerなどのマルコフ鎖モンテカルロ法である。しかし、最近では統計学者によるオリジナルなアルゴリズムも提案されてきており、常識となりつつある。メトロポリス法などの統計物理の手法を統計的情報処理に応用した先駆的な仕事のひとつとして、OgataとTanemuraのもの(Ogata and Tanemura(1981,1985)、Ogata(1990))がある。

1. 転送行列・転送積分

   1次元モデル(たとえば時系列モデル)では、事後分布での期待値は転送積分法で計算できる。統計の分野ではこれは非ガウスフィルター(Kitagawa(1987))として知られている(ガウス的な場合が、いわゆるカルマンフィルターである)。音声認識などでよく使われる'隠れマルコフ鎖に対するBaumのアルゴリズム'も本質的に同じものである(たとえばDevijver(1987)参照)。画像処理への応用については、Derin et al.(1984)やDevijver(1987)がある。類似の(recursiveな)方法で、DNAのアラインメントを"有限温度的"に行なった例として(Thorne et al.(1991,1992))がある。この例では、アラインメントを"温度零"で行なった場合に、推定された分岐年代にバイアスが生じる可能性が示唆されている点が興味深い(分岐年代がhyper parameterに相当すると考えられる)。Rujánは符号解読におけるViterbi decordingの有限温度版が転送行列法に相当すると述べているが、一般に、dynamic programmingといわれるものが転送行列・転送積分法の"絶対零度版"であることがしばしば観察される。

2. マルコフ鎖モンテカルロ法

   空間モデル・ネットワークモデルでは、転送行列・転送積分法の利用には限界があるので、マルコフ鎖モンテカルロ法の適用が考えられる。しかし、緩和が遅いため、うまくいかないことも多い。この場合、local minimumを避ける手段としてのsimulated annealingは役に立たない(目的が最適化ではないから)。おそらく、multicanonical algorithmやその変形が有望と考えられる。筆者も実験中であるが、簡単な場合以外は、まだうまくいっていない。これらの方法については、次の節(5.2節)で論じる。ベイズ統計へのマルコフ鎖モンテカルロ法の応用についての総説は、たとえば、Journal of Royal statistical society Ser.B (1993) 55 No.1 pp.3-102. を参照されたい(過去数年の論文の数に注目!)。画像関係については、IEEE Patern Analysis and Machine Intelligenceのバックナンバーを調べるのも有益かも知れない。

3. 平均場近似

   平均場近似をBoltzmann Machineの学習に応用した仕事として、Peterson and Hartman(1989)が、画像へ応用した仕事として、Geiger and Girosi(1991)、Zhang(1992)などがある。これらの論文のなかには、平均場近似の有限温度での期待値をもとめるアルゴリズムとしての側面も述べられている。ただし、そういう点が、どこまで読者に理解されているかはわからない。平均場近似と同じ近似を独自に考案し、ベイズ的な画像処理の問題に適用した仕事として、Kay and Titterington(1986)がある。この発展として、フィルタリングの手法と組み合わせて、クラスター平均場近似のようにすることも試みられているようである。

   イジング模型による2次元画像の再構成についていえば、平均場近似の精度は満足とはいえない。とくに、marginal likelihoodを用いたhyper parameterの推定には困難が大きい(Iba, unpublished note[2])。

4. 繰り込み群

   繰り込み群を画像処理に応用した研究としては、Gidas(1989)がある。しかし、重点は最適化にあるようで、筆者が理解した限りでは、有限温度的な側面は重視されていないように見える。また、marginal likelihoodを利用してのhyper parameterの学習は試みられていないようである。なお、最適化を階層的な方法で行なうアルゴリズムは、これ以外にも、マルチグリッド法の名称のもとに盛んに研究されている。

5.2　拡張アンサンブルの方法

必要な確率分布そのものでなくなんらかの意味で拡張ないし修正した確率分布に関してそれを不変にするマルコフ鎖を構成する方法が最近注目されている。これらの方法では、simulated annealing と違って、事後分布からのサンプルを作りだすことができる。

以下の各手法は、それぞれ独立の発見であるが、基本原理には共通のものがある。

1. "Multicanonical" algorithm (Berg and Neuhaus(1992), Berg and Celik(1992))
2. "Simulated tempering" algorithm (Marinari and Parisi(1992))
3. - "時間的一様な並列アニーリング" algorithm (Kimura and Taki (1990))
   - "Metropolis-coupled chains" algorithm (Geyer(1991))
   - "不公平な genetic" algorithm (Takahata(1992))

これらの手法を用いれば、通常のメトロポリス法などと同様に、目的とする確率分布からのサンプルを生成することができる。この点が、最適化のみを目的とする Simulated Annealing 法とは大きく異なるところである。

ここでは、最後の群 (3) の手法についてだけ説明する。筆者はこれを前のふたつを参考にして思いついたのだが、非常に多くの人が独立に同じ方法を考えていることがわかって、ちょっと驚いている。ただ、このうちベイズ統計との関連をはっきり意識しているのは、おそらく "Metropolis-coupled chains" algorithm だけで、あとの人たちは最適化手法としてとらえていると思われる (詳細釣合の条件については言及しているが、できた分布の利用法は考えていないように思われる)。なお、Takahata のものは、温度の違うもの同志の recombination を考えている点で他と異なっている。

この手法では、条件のことなる確率分布 $P_i(i \in \{1,..,M\})$ を生成するマルコフ鎖を複数個同時にシミュレートしながら、その間の状態の '入れ替え' を確率的に行なう。入れ替わりが十分に頻繁におこり、P_i の中に短い緩和時間でシミュレートできるものが含まれれば、全体の緩和が速くなることが期待できる。'入れ替え' の際には同時分布関数

$$\tilde{P}(x_1, x_2, ..., x_M) = P_1(x_1)\, P_2(x_2) .... P_M(x_M) \tag{36}$$

についていわゆる詳細釣合の条件が満たされるようにすれば、マルコフ鎖モンテカルロ法として必要な性質が満たされる。

たとえば、条件のことなる確率分布として温度 T_m の違うギブス分布の族 $\{P_m(x)\}$

$$P_m(x) = \frac{\exp(-\frac{E(x)}{T_m})}{Z_m} \tag{37}$$

を考えた場合には、同時分布関数を不変に保つ入れ替えのアルゴリズムとして、

1. $E(x_{m+1})$, $E(x_m)$ を計算する。
2. $\triangle = -(\,E(x_{m+1}) - E(x_m)\,)(\,\frac{1}{T_{m+1}} - \frac{1}{T_m}\,)$ を計算する。
3. 乱数 $rnd \in [0,1]$ を発生させ、$rnd < exp(-\triangle)$ なら x_m と x_{m+1} を "入れ替える"。
   すなわち、いままで、温度 T_m で動いていた系の状態を初期状態として温度 T_{m+1} の計算をはじめ、温度 T_{m+1} で動いていた系の状態を初期状態として温度 T_m の計算をはじめる。

という方法が考えられる。各温度 T_m の系はそれぞれ独立にメトロポリス法もしくは熱浴法によって時間発展するとし、それに各 m についての上述の操作を組み合わせたものが、全体のアルゴリズムを構成する。各 m について T_m と T_{m+1} が十分近ければ、入れ替わりが頻繁に起こり、高温での速い緩和が、低温での緩和を促進する効果を持つことが期待される。

“曖昧さを含む分類の問題”の簡単な場合について、このアルゴリズムを実験したところ非常にうまくいくことが示された(Iba(1993)、そのころはまだオリジナルのアルゴリズムだと思っていた)。しかし、もっとlocal minimaの影響の厳しい、“Caucy分布による平滑化・変化点検出”に応用したところうまくいかないことがわかり、現在検討中である。

付録

Nishimori(1981)の議論の一部にベイズ統計の観点からの再解釈を与えるのがこの付録の趣旨である。この論文で導かれている式のうち、(1)Nishimori line上でのエネルギーの等式、(2)比熱の不等式を考え、これらが、marginal likelihoodを使ってhyper parameterを推定する問題から自然に導かれることを示す。

いま、likelihood $L(y|x)$ がhyper parameter α を含むとして、x について周辺化したlikelihood

$$l(\alpha)=\sum_x L_\alpha(y|x)\pi(x) \tag{38}$$

を考える。α の真の値を α_0 とし、量 $A(y)$ について、

$$[A(y)]_y=\sum_y A(y)L_{\alpha_0}(y|\epsilon)\pi(\epsilon) \tag{39}$$

$$[[A(y)]_y]_\epsilon=\sum_\epsilon\sum_y A(y)L_{\alpha_0}(y|\epsilon)\pi(\epsilon) \tag{40}$$

と定義する。marginal likelihoodを使った推定がうまくいくためには、marginal likelihoodの期待値が α の真の値 α_0 のところで最大値をとることが期待される。すなわち、任意の α について、

$$[[\log l(\alpha_0)]_y]_\epsilon \geq [[\log l(\alpha)]_y]_\epsilon \tag{41}$$

とならなくてはならない。この式は情報量に関してよく知られた不等式

$$\sum_y Q(y)\log\frac{q(y)}{Q(y)}\leq 0 \tag{42}$$

からすぐ示すことができる。ただし、$Q(y)$ および $q(y)$ は y 上の任意の確率分布である。

これからただちに、

$$[[\frac{d}{d\alpha}\log l(\alpha)|_{\alpha=\alpha_0}]_y]]_\epsilon=0 \tag{43}$$

$$[\frac{d^2}{d^2\alpha}\log l(\alpha)|_{\alpha=\alpha_0}]_y]_\epsilon\leq 0 \tag{44}$$

となることが出てくる。2番目の式(44)は極大条件である。これらの式は、直接計算して示すこともできる(その方がNishimoriの原証明により近くなる)。

hyper parameterとして K を考え、$\pi(x)$ を $\{\pm 1\}$ 上の一様分布とし、

$$L(\{J_{ij}\}|\{\sigma_i\})=\frac{\exp(K\sum_{(ij)}J_{ij}\sigma_i\sigma_j)}{(2\cosh K)^n} \tag{45}$$

とおく。n は (ij) 対の数である。また $J_{ij}=\pm 1$、$\sigma_i=\pm 1$。このlikelihoodについて(43),(44)を考えるが、この種の場合の特別な事情として、gauge invarianceから、ϵ に関する平均を取り去って、ϵ を適当な状態、たとえばferromagnetic stateに固定することができる。この場合、(45)より、$\{J_{ij}\}$ について

の平均 $[\,]_y$ は分布

$$P(\{J_{ij}\}) = \Pi_{(ij)}\{p\delta(J_{ij}-1)+(1-p)\delta(J_{ij}+1)\} \tag{46}$$

について平均することに帰着される。これを $[\,]_p$ で示そう。ここで、

$$p = \frac{\exp(K)}{\exp(K)+\exp(-K)} \tag{47}$$

である。E を

$$E = -\sum_{(ij)} J_{ij}\sigma_i\sigma_j \tag{48}$$

と定義し、また、

$$\langle E\rangle_K = \frac{\sum_{config.} E\exp(K\sum_{(ij)} J_{ij}\sigma_i\sigma_j)}{\sum_{config.}\exp(K\sum_{(ij)} J_{ij}\sigma_i\sigma_j)} \tag{49}$$

などと定義するとき、(43)(44) は、それぞれ、

$$[\langle E\rangle_K]_p = -n\tanh K \tag{50}$$

$$[\langle E^2\rangle_K - \langle E\rangle_K{}^2]_p \leq \frac{n}{\cosh^2 K} \tag{51}$$

となる。(46),(47) のもとでこれらが成立するということが、Nishimori(1981) で導かれたエネルギーの等式と比熱の不等式である。

参考文献

(文献が入手困難な場合は、e-mail と住所を明記の上、筆者までお問い合わせください。)

統計学におけるマルコフ鎖モンテカルロ法についての討論つきの特集:
Journal of Royal statistical society Ser.B (1993) 55 No.1 pp.3-102.

以下はABC順。

Akaike,H.(1980)
Likelihood and Bayes procedure
Bayesian statistics
Eds. Bernardo,J.M, DeGroot,M.H., Lindley,D.V., and Smith,A.F.M.
University press Valencia.

Berg,B.A. and Neuhaus,T.(1992)
Multicanonical ensemble: a new approach to simulate first-order phase transitions
Phys.Rev.Lett. 68 9-12.

Berg,B.A. and Celik,T.(1992)
New approach to spin-glass simulations
Phys.Rev.Lett. 69 2292-2295.

Bialek,W. and Zee,A.(1987)
Statistical mechanics and invariant perception
Phys.Rev.Lett. 58 741-744.

Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977)
Maximum likelihood from incomplete data via the EM algorithm
The Journal of the Royal Statistical Society Ser.B 39 1-38 (with discussions).

Derin,H., Elliott,H., Cristi,R. and Geman,D.(1984)
Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields
IEEE Transactions on Pattern Analysis and Machine Intelligence 6 707-720.

Devijver,P.A. and Dekesel,M.M. (1987)
Learning the parameters of a hidden Markov random field image model: A simple example
NATO ASI series F30
Pattern Recognition Theory and Applications
Ed. P.A.Devijver and J.Kittler
Springer-Verlag.

Geiger,D. and Girosi,F.(1991)
Parallel and deterministic algorithms for MRF's : surface reconstruction and integration
IEEE Transactions on Pattern Analysis and Machine Intelligence 12 401-412.

Geman,S. and Geman,D.(1984)
Stochastic relaxation, Gibbs distributions,
and the Bayesian restoration of images

IEEE Transactions on Pattern Analysis and Machine Intelligence 6 721-741.

Geman,S. and McClure,D.E.(1987)
Statistical methods for tomographic image reconstruction
Proc. of the 46th Session of the ISI, Bulletin of the ISI, Vol.52.

Geyer,C.J.(1991)
Markov chain Monte Carlo maximum likelihood
In Computing Science and Statistics:
Proc. of the 23rd symposium on the interface(E.M.Keramides,ed.).

Gidas,B (1989)
A renormalization group approach to image processing problems
IEEE Transactions on Pattern Analysis and Machine Intelligence 11 164-180.

Good,I.J.(1965)
The estimation of probabilities
MIT press Cambridge .

Hinton,G.E. and Sejnowski,T.J.(1986)
Learning and relearning in Boltzmann machines
in Parallel Distributed Processing Vol.1
Eds. Rumelhart,E. and McClelland,J.L.
MIT press Cambridge .

Iba,Y.(1989)
Bayesian statistics and statistical mechanics
In cooperative dynamics in complex physical systems
Ed. Takayama,H. Springer-Verlag Berlin .

Iba,Y.(1991a)
Macroscopic parameter estimation from incomplete data with Metropolis-type Monte Carlo algorithm
Proceedings of the institute of statistical mathematics Vol.39 No.1 1-21
(in Japanese with English summary).

Iba,Y.(1991b)
帰無仮説下での周辺ゆう度の振舞いとパラメータ空間の次元
統計数理研究所 年度末発表会 (1991 年度)
要旨
Proceedings of the institute of statistical mathematics Vol.39 No.1 153-156
(in Japanese).

Iba,Y.(1991c)
Metropolis-type Monte Carlo algorithm and
quasi Bayesian estimation procedure
Proceedings of the institute of statistical mathematics Vol.39 No.2 225-244
(in Japanese with English summary).

Iba,Y(1992)
An application of Metropolis-type algorithm to a complex classification problem
ISM Research Memorandum No.440.

Iba,Y(1993) 統計数理研究所 年度末発表会 (1992 年度)
統計数理 (Proceedings of the institute of statistical mathematics) に要旨掲載予定

Kay.J.W. and Titterington D.M. (1986)
Image labelling and statistical analysis of incomplete data
Proc. 2nd. Int. Conf. Image processing and Applications
Conf. Publ. No.265 London Inst. Elec. Engrs. 44-48.

Kimura,H. and Taki,K (1990)
On a time-homogeneous parallel annealing algorithm
電子情報通信学会 NC-90-1 1-8
(in Japanese).

Kitagawa,G. and Gersch,W. (1985)
A smoothness priors time-varying AR coefficient modeling of nonstationary covariance time series
IEEE Transactions on automatic control AC-30 48-56.

Kitagawa,G.(1987)
Non-Gaussian state space modeling of nonstationary time series (with discussion)
Journal of the American Statistical Association 79 1032-1063.

MacKay,D.J.C.(1992)
Neural Computation 4 415-447; 448-472.

Marinari,E. and Parisi,G.(1992)
Simulated Tempering: a New Monte Carlo Scheme
Europhys.Lett. 19 451-458.

Marroquin,J.(1985)
Optimal Bayesian estimators for
image segmentation and surface reconstruction
MIT A.I. Memo. 839.

Marroquin,J., Mitter,S. and Poggio,T. (1987)
Probabilistic solution of ill-posed problems in computational vision
Journal of the American Statistical Association Vol.82 76-89.

Nishimori,H. (1981)
Internal energy, specific heat and correlation function of the bond-random Ising model
Progress of Theoretical physics 66 1169-1181.

Nishimori,H. (1993)
Optimum decoding temperature for error-correcting codes
Journal of Physical Society of Japan 62 2973-2975.

Ogata,Y.(1990)
A Monte Carlo method for objective Bayesian procedure
Annals of Institute of Statistical Mathematics 42 403-433.

Ogata,Y. and Tanemura,M.(1981)
Estimation of interaction potentials of spatial point patterns through maximum likelihood procedure
Annals of Institute of Statistical Mathematics 33 315-338.

Ogata,Y. and Tanemura,M.(1984)
Likelihood analysis of spatial point patterns
Journal of Royal statistical society Ser.B 46 496-518.

Peterson,C. and Hartman,E. (1989)
Explorations of the mean field thoery learning algorithm
Neural Networks 2 475-494.

Ruján,P.(1993)
Finite temperature error-correction codes
Phys.Rev.Lett. 2968-2971.

Sakamoto,Y.(1985)
カテゴリカルデータのモデル分析
共立出版　(英語版もあり)

Soulas,N.(1993)
Spin-glasses,error-correcting codes and finite temperature decoding
preprint.

高畠一哉 (1992)
Unfair genetic algorithm
1992年 電子情報通信学会秋期大会 A-172.

Tanabe,K.(1985)
ベイズモデルと ABIC
オペレーションズ・リサーチ 1985年3月号 178-183.

Thorne,J.L., Kishino,H. and Felsenstein,J. (1991)
An evolutionary model for maximum likelihood alignment of DNA sequences
Journal of molecular evolution 33 114-124.

Thorne,J.L., Kishino,H. and Felsenstein,J. (1992)
Inching toward reality; an improved likelihood model of sequence evolution
Journal of molecular evolution 34 3-16 (1992).

Yanagimoto,T and Yanagimoto,M(1987)
The use of marginal likelihood for a diagnostic test for the goodness of fit of the simple linear regression model

Technometrics 29 95-101.

Wahba,G(1990)
Spline Models for observational data
SIAM, Philadelphia Pennsylvania.

Zee,A.(1987)
Some quantitative issues in the theory of perception
preprint(基礎物理学研究所 preprint library にあった。本論文不明。)

Zhang,J.(1992)
The mean field theory in EM procedures for Markov random fields
IEEE Transactions on signal processing 40 2570-2583.

Unpublished Notes

[0] は私家版で配布したレビューであるが、著書に引用してくださった方がいるため、入手困難であるとのお叱りをうけた。大部古くなってしまったが、近く出版できるように努力中である。[1] と [2] はその後の急速な発展によって、出版する価値が減少してしまった。[3] は自分でも良く理解できない (?) 変な内容のもの。

[0] 統計物理と統計的情報処理 1990/2/13

[1] ベイズ的画像処理における事後分布と最ゆう解 1989/06/20

[2] 画像処理における平均場近似 1989/06/20

[3] 情報理論における Gallager formalism とレプリカ法 1989/6/28

文献追加

マルコフ場の情報処理への応用については、次の本がある。

Markov Random Fields: Theory and Application
Eds. Chellappa,R. and Jain,A.
Academic Press San Diego 1993.